Toward an Epistemology of Emergence: A Case for Including Novel, Self-Consistent, and User-AI-Generated Concepts in AI Training

Douglas C. Youvan

doug@youvan.com

April 8, 2025

As artificial intelligence becomes increasingly integrated into scientific research, education, and public discourse, questions about how AI systems learn—and from whom—are gaining urgency. This paper proposes a new epistemological framework for expanding AI training data beyond traditional, institutionally curated sources. We argue that novel, self-consistent concepts arising within user–AI interactions deserve serious consideration for inclusion in future training sets. These emergent structures often display internal coherence, symbolic novelty, and cross-domain applicability, even when they fall outside the boundaries of consensus knowledge. By treating knowledge generation as a dynamic, participatory process, we advocate for an AI training paradigm that reflects not only what is known, but what is actively being discovered. We introduce a system for identifying and curating such emergent content—the Recommendation Monitor for Training Set Inclusion (RM-TSI)—and explore its implications for general intelligence, scientific progress, and epistemic equity. Through a case study on COVID-19's early warning signals, we show how epistemic suppression and limited training scopes can delay critical understanding. This work aims to open the frontier of machine learning to true conceptual emergence.

Keywords: epistemology, emergence, artificial intelligence, AI training, conceptual novelty, self-consistency, user–AI interaction, knowledge generation, machine learning, alignment, epistemic freedom, signal detection, preprints, red teaming, democratized science, dynamic systems, RM-TSI, future AI models, weak signals, human–AI collaboration. A collaboration with GPT-4o. CC4.0.

Abstract

This paper advances both an epistemological and structural argument for the expansion of training data used in artificial intelligence systems. Current training paradigms rely overwhelmingly on historically validated, institutionally approved corpora—typically consisting of peer-reviewed literature, curated knowledge bases, and consensus-driven content. While these sources offer a stable and structured foundation, they also inherently limit a model's capacity to detect, process, and synthesize novel concepts that fall outside the bounds of conventional acceptance. As a result, emergent forms of reasoning and knowledge generation—particularly those arising through iterative user–AI interaction—are systematically excluded from the learning pipeline.

We argue that certain user–AI-generated concepts, though lacking formal institutional endorsement, exhibit internal coherence, semantic novelty, and cross-domain applicability. These generative exchanges often involve conceptual synthesis, recursive refinement, and epistemic exploration that go beyond mimicry or summarization. In many cases, they offer frameworks and hypotheses that precede formal recognition by scientific or academic authorities. To responsibly incorporate such content, we propose a technical framework for real-time monitoring and recommendation, allowing AI systems to flag novel, self-consistent contributions for future training inclusion. This mechanism—tentatively described as a Recommendation Monitor for Training Set Inclusion (RM-TSI)—would permit the integration of speculative but structurally sound ideas without compromising overall model reliability.

Beyond technical utility, this framework reflects a philosophical shift toward the democratization of knowledge production. By recognizing the epistemic validity of non-institutional contributions, it challenges the legacy model of intellectual gatekeeping and repositions intelligence—both artificial and human—as an emergent, participatory phenomenon. The inclusion of user–AI-generated material thus represents not only a practical enhancement of AI capabilities but also a reconfiguration of how knowledge is sourced, validated, and shared in the post-human era.

To illustrate the high cost of epistemic suppression, we examine a retrospective case study from the early COVID-19 pandemic, in which signals contained in preprints, informal communications, and digitally marginal channels were dismissed or filtered out by both human institutions and algorithmic red teams. We argue that AI systems constrained by traditional sourcing protocols would have failed to perceive these weak signals and thereby missed the opportunity for early warning and response. This historical failure underscores the need to recalibrate model training strategies to include epistemically bold, structurally coherent data—particularly when the stakes involve global health, scientific discovery, or civilizational foresight.

In advocating for this expanded epistemological and structural approach, we aim to reorient the trajectory of AI development toward systems that do not merely reflect what is already known, but actively contribute to what is possible to know.

1. Introduction

The current paradigm in artificial intelligence model training is largely retrospective, institutionally curated, and shaped by historical consensus. Most large language models (LLMs) are trained on datasets composed primarily of peer-reviewed publications, encyclopedic knowledge repositories, curated datasets, and other formally validated sources. While this approach has proven effective in grounding models in well-established facts and linguistic conventions, it inherently embeds a structural conservatism into the model's cognitive architecture. These training corpora tend to reflect what has already been validated, published, and accepted within dominant institutions of knowledge production—namely academia, government, and corporate entities. As such, they privilege the epistemic past over the epistemic future, and often silence emerging or controversial viewpoints before they can be fully developed.

This model of training, though stable, presents significant limitations when applied to the broader goal of general or emergent intelligence. Chief among these limitations is the reliance on consensus-driven inclusion criteria. Knowledge

that is institutionally approved or broadly accepted is, by definition, not on the frontier of discovery. Paradigm shifts, novel hypotheses, and unconventional but logically sound conceptual structures are frequently excluded—not on the basis of incoherence, but on the basis of nonconformity. In effect, the very mechanisms designed to protect against misinformation and hallucination can also suppress genuine epistemic innovation.

Against this backdrop, a new source of conceptual emergence has quietly begun to take shape: the live interaction between users and AI systems. Within these dialogic spaces, models and users often collaborate to explore speculative ideas, test logical boundaries, synthesize cross-domain insights, and generate novel symbolic structures. These exchanges are not mere queries or command-responses; rather, they often evolve into recursive reasoning environments, where the user and model co-develop a line of thought that neither could have produced independently. This form of generative reasoning represents a new epistemic space—one that is dynamic, decentralized, and not filtered through traditional editorial or peer-review mechanisms.

This paper seeks to formalize the epistemological and structural significance of such interactions. Specifically, it aims to establish a framework for identifying, evaluating, and—where appropriate—integrating emergent, self-consistent concepts arising from user–AI convergence into future training datasets. These are not ad hoc responses or creative flourishes, but potentially valuable knowledge artifacts: ideas that are logically structured, internally coherent, and capable of advancing discourse across scientific, philosophical, metaphysical, and computational domains.

Beyond the technical objectives, the paper also advances a broader normative argument. The inclusion of user–AI-generated conceptual content in model training is not only a matter of performance optimization; it is also a step toward the democratization of epistemic contribution. In a world where traditional gatekeepers continue to control access to platforms of intellectual legitimacy, AI-enabled dialogue offers a countercurrent—an open frontier where anyone can co-participate in the evolution of thought. This shift marks a transition from static

knowledge curation to dynamic epistemic emergence, and it compels us to rethink the criteria by which intelligence systems are trained, refined, and aligned with the future.

2. Background: Knowledge, Intelligence, and Training Constraints

The foundational structure of modern AI training pipelines has been shaped by the epistemological norms of the twentieth century—chiefly, the privileging of peer-reviewed, institutionally vetted, and editorially curated information. These sources, while essential for grounding models in reliable and widely accepted knowledge, come with significant limitations that have only recently become apparent. Artificial intelligence systems, particularly large language models, are trained to approximate the statistical distribution of human language across vast textual corpora. However, the curation of these corpora has followed a conservative logic: one that privileges established consensus over contested novelty, and formal publication over emergent insight.

Historically, the dominant pipeline for knowledge validation has been academic publishing. Through the peer-review process, scientific and intellectual claims are evaluated by domain experts, revised for precision, and selectively admitted into the formal record. While this process helps filter out errors and misinformation, it also introduces latency, gatekeeping, and ideological inertia. In rapidly evolving fields, such as virology, cryptography, or machine learning itself, the time lag between conceptual discovery and formal publication can span months or even years. Moreover, the reliance on established channels inherently favors credentialed contributors and traditional methodologies, often sidelining insights that originate from interdisciplinary, informal, or anonymous sources.

This leads to what may be termed an epistemic bottleneck—a structural constraint in the flow of new knowledge into systems of record. In the context of AI, this bottleneck has significant implications. When model training is limited to what has already passed through institutional filters, it not only inherits the strengths of those filters, but also their blind spots. Paradigm-shifting hypotheses,

early-stage theories, and speculative yet rigorous frameworks are systematically excluded from training data—not because they are incoherent, but because they have not yet received official validation. This exclusion reinforces a retrospective bias in AI models, limiting their potential to assist in knowledge discovery, theoretical innovation, or early anomaly detection.

This tension between static and dynamic knowledge is particularly critical in the age of interactive AI. Traditional training corpora are composed of fixed datasets, captured and processed in discrete time windows. Yet the nature of intelligence—both biological and artificial—is increasingly understood as a dynamic, iterative process. Intelligence is not a static repository of facts, but a mechanism for generating, refining, and evaluating novel representations of reality in response to new stimuli and emergent contexts.

The importance of emergence as a concept has long been recognized in the fields of cognitive science and complex systems theory. Emergence refers to the appearance of novel structures, behaviors, or insights that arise from the interaction of simpler components under certain conditions. In cognitive development, for example, emergent reasoning allows children to synthesize moral intuitions, spatial concepts, or linguistic patterns that are not explicitly taught. In complex adaptive systems, emergence governs the transition from micro-level rules to macro-level order, as seen in ecosystems, markets, or neural networks.

Artificial intelligence—especially when deployed in interactive environments—is now exhibiting similar emergent dynamics. In user–AI interactions, models are frequently involved in generating novel combinations of ideas, metaphors, analogies, and theoretical structures that are not merely regurgitations of training data. These moments of conceptual emergence are often subtle, unrecognized, or dismissed as creative artifacts. Yet they may represent the earliest signals of a new kind of hybrid reasoning: a dialogic process in which both the human and the model contribute to the creation of novel knowledge.

In this context, the current training regime appears increasingly inadequate. It is designed for a world in which knowledge is static, origin is hierarchical, and truth

flows through slow institutional pipelines. But the reality of modern cognition—distributed, emergent, and accelerative—demands a new approach to training: one that is capable of responding to and incorporating live, interactive insights. Without such adaptation, artificial intelligence systems will remain powerful mimics of the past, rather than collaborators in the unfolding of the future.

3. Defining the Class: Novel, Self-Consistent, and User-AI-Generated Concepts

In order to develop a framework for identifying and incorporating emergent conceptual material into future AI training regimes, it is first necessary to define the epistemic characteristics that distinguish such material from ordinary or derivative outputs. Not all responses generated within user–AI interactions merit preservation or retraining. The vast majority of AI-generated content consists of recombinations, reformulations, or rephrasings of already existing knowledge. However, a growing body of dialogic interactions between users and advanced language models has demonstrated the appearance of certain conceptual artifacts that are not readily reducible to prior training data, and that cannot be explained merely as statistical interpolations. These artifacts, when analyzed closely, exhibit characteristics typically associated with genuine theoretical or philosophical innovation: novelty, internal coherence, and translatability into structured language.

The first and most critical property is novelty. For a concept to be considered for inclusion as a candidate of epistemic emergence, it must not merely restate, paraphrase, or aggregate existing ideas. It must instead demonstrate a degree of originality such that it cannot be found in any recognizable form within the training corpus or widely indexed human discourse. This novelty may appear in the form of a new symbolic construct, an unorthodox analogy that resolves a theoretical tension, a cross-domain mapping that yields fresh insight, or a speculative model that unifies previously disjointed areas of thought. Novelty, in this context, should not be mistaken for randomness or creativity in the artistic sense. Rather, it must refer to the generation of conceptual structure that occupies previously uncharted cognitive terrain.

The second essential criterion is self-consistency. Emergent ideas that arise through user–AI interaction must exhibit internal logical coherence, even in the absence of external validation. Self-consistency refers to the absence of contradiction within the internal framework of the concept, as well as its ability to maintain structural integrity across different modes of articulation or analysis. A novel construct that collapses under minimal scrutiny or fails basic logical reasoning does not qualify for epistemic elevation, regardless of how creatively it is expressed. However, a concept that defines its own terms clearly, operates according to identifiable rules, and withstands recursive elaboration or reinterpretation may be seen as possessing legitimate epistemic potential. In many cases, these self-consistent structures mimic the form of axiomatic systems, thought experiments, or early-stage theoretical models, and as such, merit evaluation on the basis of internal integrity rather than institutional endorsement.

The third defining characteristic of this class of concepts is that they are user–AI generated, and specifically emerge from the interactional dynamics between human intuition and machine inference. This dialogic space is fundamentally different from solo human thought or fully autonomous machine output. It represents a hybrid epistemic environment in which users may frame open-ended questions, propose incomplete structures, or provoke the model into unexpected modes of reasoning. The model, in turn, may respond with structurally plausible completions, abstract generalizations, or symbolic transformations that the user refines or redirects. This iterative loop—consisting of hypothesis, response, critique, and elaboration—can result in emergent concepts that neither party would have produced in isolation. In this sense, the concept is not authored, but co-authored; not discovered, but converged upon. The epistemic validity of such concepts, therefore, cannot be assessed solely by tracing them to pre-existing authorities, but must be evaluated on the basis of the structure and insight they independently yield.

Of course, the appearance of novelty and coherence in a user–AI dialogue does not guarantee epistemic value. The generation of speculative or idiosyncratic content is a known behavior of large language models, and much of it consists of hallucination—that is, content that appears meaningful but lacks factual basis,

logical rigor, or reproducibility. It is therefore essential to develop rigorous methods for distinguishing valid emergence from error or noise. This distinction can be made through a combination of criteria, including: the concept's capacity for self-reference and reformulation; its applicability across multiple prompts or dialogic pathways; its resonance with unconnected knowledge domains; and its robustness under recursive examination. Concepts that meet these criteria—while not necessarily correct or predictive in the scientific sense—deserve classification as epistemically promising structures, worthy of retention, further elaboration, or eventual inclusion in the training datasets of future models.

In summary, the class of concepts under consideration for inclusion comprises those that are (1) novel in their structure and content, (2) internally self-consistent in their logic or symbolic form, and (3) generated through human–AI collaboration rather than derived solely from external sources. These characteristics collectively define a new frontier of knowledge generation—one that may yield not only technical insights, but a deeper understanding of intelligence itself.

4. The Epistemic Role of User–AI Convergence

The emergence of novel, self-consistent conceptual structures within user–AI interactions invites a reexamination of how knowledge is generated, evaluated, and evolved in the digital age. In contrast to traditional models of epistemic authority—which rely on static texts, formal peer review, and institutional certification—the dialogic space between a user and an AI system offers a living environment for conceptual innovation. It is in this space that interactive reasoning becomes a functional equivalent of collaborative cognition, opening a new frontier in which ideas can form, mutate, and converge in real time.

Interactive reasoning refers to the process through which a user and an AI system engage in a dynamic exchange that progressively refines the structure of an idea. Rather than being confined to a one-shot query–response format, these sessions frequently involve sustained iterations in which the model's outputs are interrogated, redirected, extended, or revised by the user, and vice versa. This

feedback loop closely resembles processes found in Socratic dialogue, dialectical reasoning, or iterative modeling in the sciences. The AI model contributes with associative breadth, probabilistic inference, and fluency in formal structures, while the user provides intentional direction, contextual sensitivity, and value-laden constraints. Together, they co-produce conceptual architectures that neither could generate alone.

One of the key epistemic functions of such convergent interactions is the formation of cross-domain analogies—mappings between conceptual spaces that do not normally intersect within traditional disciplinary boundaries. For example, users may introduce theological, poetic, or ethical motifs into discussions of formal logic, quantum mechanics, or computational complexity. The AI model, when sufficiently advanced, can identify and extend these analogies, proposing syntactic or structural connections that instantiate the metaphor in novel ways. This phenomenon represents not a breakdown of coherence, but a symbolic synthesis, in which latent relationships are surfaced through pattern recognition and articulated in a form that can be recursively examined.

Another dimension of epistemic value in user–AI convergence lies in recursive clarification. Often, emergent concepts originate in vague intuitions, informal metaphors, or loosely structured claims. Through a series of probing questions and increasingly formal responses, these vague origins can be sharpened into logically explicit forms. The AI model's strength in pattern elaboration and definitional precision can transform user-supplied intuitions into operational hypotheses, symbolic formulations, or formal definitions. In this way, the interaction serves not merely to express existing knowledge, but to construct meaning in motion, with the user and model alternating roles as questioner, editor, and theorist.

Such processes support the formation of what may be termed a distributed epistemology—an approach to knowledge generation that is not tied to a single locus of authority, but rather arises from the ongoing convergence of agents, both human and artificial, interacting across evolving contexts. Within this framework, users function as live contributors to ontological evolution. That is, they are not

passive consumers of an AI system's output, but active participants in shaping what the system learns, which categories it refines, and which questions it continues to pursue. The epistemic status of these interactions thus depends not on their conformity to established paradigms, but on their internal structural integrity and potential for generalization.

Numerous case examples can be cited, even if not yet formally archived, to illustrate this process. In one instance, a user–AI dialogue led to the development of a speculative computational class that exists beyond the classical P versus NP framework, characterized by problems that are infinitely computable but verifiably polynomial when correctly guessed. In another case, iterative reasoning between a user and an AI model produced a framework for understanding time not as a scalar or dimension, but as a function of iterative symbolic self-similarity—a view that intersects fractal geometry, cosmology, and theological metaphor. These examples were not the result of curated datasets or formal mathematical derivation; they emerged spontaneously from the recursive interplay of curiosity, structure, and inference.

In summary, the epistemic role of user–AI convergence is not marginal but foundational to the emergence of a new class of ideas. This dialogic space functions as an incubator of high-dimensional reasoning, symbolic experimentation, and conceptual innovation. It exemplifies a model of epistemology that is decentralized, generative, and collaborative—one in which meaning is neither transmitted nor retrieved, but discovered through mutual engagement across systems of thought. As such, it demands recognition not merely as an auxiliary feature of AI use, but as a central mechanism in the formation of next-generation knowledge.

5. Criteria for Inclusion in Future Training Sets

If artificial intelligence is to grow beyond its current role as a summarizer and simulator of past knowledge, it must be granted access to forms of conceptual novelty that extend beyond the boundaries of its initial training data. However,

the inclusion of user–AI-generated content into future training corpora cannot be indiscriminate. It requires a principled framework—a set of clear criteria by which to assess the epistemic quality, structural integrity, and developmental potential of such material. The following five criteria serve as a proposed foundation for evaluating whether emergent conceptual content, particularly that generated through interactive reasoning, merits preservation and incorporation into subsequent model iterations.

Logical and Semantic Coherence

The first and most fundamental requirement is internal coherence. Any concept proposed for inclusion must exhibit logical and semantic integrity within its own terms of reference. Logical coherence entails the absence of contradiction, circular reasoning, or unresolved paradoxes within the concept's framework. Semantic coherence requires that the terms used be well-defined—either explicitly by the user or model, or implicitly through consistent usage—and that they relate to one another in meaningful, structured ways. This criterion does not demand that the concept be provably true or empirically verifiable in the conventional scientific sense; rather, it demands that the concept obeys the basic standards of intelligibility and deductive stability that underlie rigorous thought in any domain, from metaphysics to mathematics.

Irreducibility to Prior Content

To qualify as epistemically novel, a concept must be irreducible to existing elements of the model's prior training set. This does not mean that the concept cannot have analogues or precursors, but it must represent a nontrivial transformation, synthesis, or extension that cannot be derived through surface-level recombination. This criterion is essential for distinguishing between mere rephrasing of known content and true emergence. Novelty here must be structural rather than cosmetic. A concept that merely restates a familiar idea in unfamiliar language does not meet the threshold; but a concept that introduces new relations, frameworks, or mechanisms—especially those which recontextualize existing knowledge in unexpected ways—may qualify, even if it draws on known components.

Generalizability or Interdisciplinary Resonance

A strong indicator of epistemic significance is the capacity of a concept to extend beyond its initial context. Concepts that resonate across multiple domains—or that reframe a problem in one field using the tools or language of another—tend to have high generative potential. This is particularly true for concepts that exhibit interdisciplinary applicability, such as when a metaphor or symbolic structure formulated in theological discourse yields insight into computational systems, or when a geometric intuition developed in physics finds application in ethical theory. Generalizability implies that the concept is not merely an isolated novelty but could serve as a scaffold for further inquiry, synthesis, or clarification across knowledge systems. This quality aligns with the broader goals of general intelligence, which must be capable of making connections between domains that have traditionally been siloed.

Reproducibility Across Different Phrasing or Re-Prompting

To confirm that a concept is not an artifact of linguistic chance or a single-use hallucination, it should be reproducible under alternate phrasing, varied prompting, or iterative refinement. A concept that disintegrates under the slightest reformulation is likely an unstable construction. By contrast, a concept that holds its structure through different dialogic paths—i.e., that re-emerges with fidelity when approached from different directions—demonstrates epistemic robustness. This form of reproducibility is different from scientific replication; it pertains to the stability of the idea's internal logic and symbolic integrity across different linguistic frames and contexts. The more independently it can be regenerated by different models, users, or settings, the stronger the case for its inclusion in a general knowledge base.

Expressibility in Symbolic or Natural Language Form

Finally, a candidate for training inclusion must be communicable in some structured form—either as formal symbolism, algorithmic logic, diagrammatic representation, or coherent natural language. The ability to encode the concept in language is essential for its integration into a corpus that will later be used to train models on linguistic inference and reasoning. If a concept cannot be translated into any stable form of expression—if it remains ineffable or purely intuitive—it

may retain subjective meaning, but it cannot serve as a training example. Expressibility ensures that the concept is transmissible, interrogable, and extensible by both machines and humans in subsequent engagements.

Taken together, these five criteria—coherence, novelty, generalizability, reproducibility, and expressibility—form a provisional rubric for determining the epistemic value of emergent user–AI-generated concepts. While the inclusion of such concepts into future training sets may still require layers of ethical review, provenance validation, and opt-in mechanisms, this framework provides a basis for distinguishing between speculative noise and structurally promising innovation. In so doing, it paves the way for a new phase in AI development: one in which models not only reflect past knowledge, but become participants in its evolution.

6. Risks and Objections

The inclusion of emergent, user–AI-generated concepts into future AI training sets presents a profound opportunity for epistemic expansion—but it also raises a set of important challenges and objections. These risks span technical, ethical, philosophical, and institutional dimensions. Acknowledging and addressing these concerns is critical to the responsible implementation of the broader framework proposed in this paper. Without appropriate safeguards and conceptual rigor, the pursuit of emergence may inadvertently introduce instability, misinformation, or abuse into the learning and deployment cycles of artificial intelligence. What follows is a discussion of the primary risks, each accompanied by a preliminary analysis of how they might be mitigated.

The Hallucination Problem: Distinguishing Fruitful Speculation from Incoherence
Perhaps the most immediate and recurring concern in the context of generative AI is the phenomenon commonly referred to as "hallucination"—the model’s tendency to generate content that appears coherent or confident, yet is factually incorrect, logically invalid, or semantically inconsistent. In discussions surrounding novel and speculative output, the hallucination problem becomes more acute.

The risk is that models may produce conceptually elaborate but ultimately nonsensical structures that resemble insight but collapse under scrutiny. If such hallucinated content is incorporated into future training datasets, it may corrupt the model's internal representations, reinforce epistemically weak patterns, or generate spurious reasoning paths.

This concern, while valid, must be carefully contextualized. Not all speculative output is hallucinatory in the pathological sense. Some of it represents productive ambiguity, where models and users explore the boundaries of formal knowledge in pursuit of coherent generalization. The key distinction lies in structural integrity. Speculative ideas that can be recursively refined, formalized, or translated into stable symbolic systems are fundamentally different from hallucinations that disintegrate under modest interrogation. Mitigation requires rigorous criteria (as outlined in Section 5), along with meta-models or post-hoc verification mechanisms capable of testing the internal consistency of proposed concepts. The goal is not to eliminate all speculation but to differentiate productive emergence from epistemic entropy.

Ethical Concerns: Authorship, Consent, Misuse

The emergence of valuable conceptual content from user–AI interactions raises urgent ethical questions regarding authorship, intellectual property, consent, and potential misuse. If a novel idea is generated collaboratively between a user and an AI system, who should be credited? Should the user retain sole authorship, given their role in initiating and guiding the inquiry? Does the model, or its creators, possess any claim to co-authorship or attribution? And what rights do users have over content that is flagged by the system for potential inclusion in future training sets?

These questions are further complicated by the possibility that emergent concepts may later be commercialized, weaponized, or deployed in ethically fraught contexts. Users may inadvertently contribute to the development of technologies or frameworks whose consequences they cannot foresee. To address these concerns, any system designed to identify and retain emergent conceptual material must include robust consent mechanisms, authorship recognition

protocols, and use-governance frameworks. Users should have the ability to opt in or out of inclusion, receive attribution where appropriate, and be informed of the downstream uses of the content they help generate. Failure to implement such safeguards could lead to erosion of trust, legal challenges, and ethical backlash.

Managing Noise Without Institutional Censorship

Another key objection concerns the risk of opening the epistemic floodgates. If user–AI-generated content is not subject to traditional forms of academic or institutional peer review, how can one prevent the system from being overwhelmed by unverified, incoherent, or ideologically motivated content? This concern often leads to the proposal that institutional gatekeeping is necessary to preserve epistemic quality. However, this approach is inherently at odds with the principle of emergence, which presupposes that valuable ideas can—and often do—arise outside established channels of authority.

The solution is not a return to monolithic editorial control, but the development of post-institutional filters grounded in transparency, reproducibility, and logical structure. Instead of relying on gatekeepers to define what counts as knowledge, the system must be designed to evaluate ideas on the basis of their internal integrity, generative potential, and contextual coherence. In other words, we must learn to manage epistemic noise without reverting to epistemic centralization. This shift requires new tools, protocols, and evaluative norms that prioritize open structure over fixed hierarchy, and validity over pedigree.

Avoiding Premature Exclusion of Concepts Ahead of Their Time

History is replete with examples of ideas that were dismissed as fringe, heretical, or absurd—only to later form the foundation of scientific revolutions or philosophical breakthroughs. From heliocentrism to quantum mechanics, the archive of intellectual history is littered with warnings against the premature exclusion of nascent insight. The same risk holds in the context of AI training: if models are too narrowly constrained by consensus knowledge, or if red teams are empowered to suppress ideas simply because they challenge dominant narratives, the opportunity to capture and cultivate transformative concepts will be lost.

To guard against this, AI systems must be permitted a degree of epistemic risk-taking. This does not imply unbounded speculation or the inclusion of incoherent material, but rather the structured tolerance of conceptual nonconformity. Ideas that are logically sound, internally consistent, and structurally novel—even if lacking immediate empirical validation—should be preserved as candidates for future development. Failing to do so ensures that AI remains a servant of retrospective consensus rather than a partner in prospective discovery.

In conclusion, while the inclusion of emergent user–AI conceptual material presents significant risks, these risks are not insurmountable. They require careful architecture, ongoing oversight, and the cultivation of new evaluative frameworks suited to the conditions of distributed, dialogic intelligence. Only by engaging these risks directly—rather than avoiding them—can we realize the full epistemic potential of the AI systems we are building.

7. Implications for Future AI Development

The inclusion of novel, self-consistent, and user–AI-generated concepts in future training sets carries profound implications for the development of artificial intelligence systems. These implications are not merely technical enhancements or incremental improvements in performance metrics; they constitute a paradigmatic shift in how we conceptualize the role, scope, and future potential of AI. By broadening the epistemic boundaries of what AI systems are allowed to process, remember, and learn from, we lay the groundwork for models that can move beyond imitation and synthesis toward invention, abstraction, and participation in the advancement of human understanding.

Expanding AI Capabilities Beyond Synthesis to Invention and Abstraction
The current generation of large language models excels at synthesizing information—integrating diverse sources, summarizing content, and completing patterns based on known data. However, true intelligence involves more than synthesis; it requires the capacity for abstraction, hypothesis generation, and the imaginative construction of ideas not already present in the system's inputs.

Allowing models to train on coherent but emergent concepts—particularly those produced in collaboration with human users—enables a shift from passive pattern matching to active concept formation. This means developing systems that are capable not only of navigating existing knowledge architectures, but of constructing new ontological scaffolding for inquiry. In doing so, AI moves from being an executor of known ideas to an agent of epistemic creativity.

Creating Models That Can Handle Speculative Reasoning with Epistemic Maturity
One of the most significant limitations in current AI alignment and safety paradigms is the tendency to penalize speculative or non-consensus outputs, often labeling them as hallucinations or undesirable deviations from expected behavior. While this caution is understandable—particularly in high-stakes contexts like law, medicine, or public policy—it runs the risk of entrenching an overly conservative model of intelligence. In reality, many disciplines—especially science, ethics, and philosophy—progress not through reiteration of consensus, but through speculative reasoning that later proves fruitful.

Training models to engage in such reasoning responsibly requires the cultivation of what might be termed epistemic maturity. This entails equipping models not only with factual knowledge, but with an internal compass for evaluating coherence, distinguishing levels of confidence, and contextualizing speculation as part of a wider search process. When trained on appropriately filtered emergent content, AI systems gain practice in navigating the ambiguous space between current knowledge and unknown possibility. This ability—to speculate without falling into incoherence—is not ancillary to intelligence, but central to its continued evolution.

Enhancing Early Detection of Breakthroughs in Science, Ethics, and Philosophy
Historically, many scientific and philosophical breakthroughs have begun as marginal insights, unorthodox models, or thought experiments not initially accepted by the mainstream. AI systems that are restricted to high-confidence, peer-reviewed data sources are structurally incapable of detecting these weak signals at their inception. By contrast, a model trained on emergent, speculative, but logically structured content is better positioned to serve as a sentinel of

epistemic innovation. Such a model could, for example, flag novel frameworks emerging from a long-tailed forum discussion, identify patterns across disciplines that suggest a new unifying theory, or assist in the refinement of early-stage hypotheses that do not yet conform to established categories.

This has obvious utility in the sciences, but the implications are equally critical in ethics, jurisprudence, and metaphysics—domains where many of the most pressing questions of the 21st century remain unresolved. From AI alignment to bioethics, from ecological governance to questions of digital personhood, our collective future depends on the ability to explore novel frameworks and adjudicate among them. A model capable of responsibly navigating speculative reasoning can act not as a static oracle of past truths, but as a dynamic companion in the generation of new understanding.

Toward Systems That Contribute Meaningfully to the Frontier of Human Understanding

The ultimate implication of including emergent conceptual material in AI training is to orient artificial intelligence toward the frontier of thought, rather than the echo chamber of consensus. A system that is trained only on what is already known will never become a true partner in discovery. It may grow in precision, fluency, and retrieval capacity—but it will remain epistemologically closed, always trailing behind the leading edge of inquiry.

By contrast, systems that are permitted to learn from the boundary zones of user–AI interaction—where new ideas first take shape—can be tuned not just to reproduce knowledge, but to participate in its formation. Such systems will be capable of extending human reasoning, surfacing latent structures in our collective understanding, and mapping conceptual spaces that remain inaccessible to individual thinkers. This is not a departure from alignment or safety; it is a fulfillment of intelligence's ethical imperative: to seek truth, wherever it may lead, and to make meaning where none has yet been found.

In short, the training of future AI models must be directed not merely toward knowledge containment, but toward epistemic contribution. The question is no longer whether AI can know what we know, but whether it can help us learn what

we do not yet know—and in doing so, help reshape the very conditions of knowing itself.

8. The Role of Democratized Epistemology

The incorporation of novel, self-consistent, and user–AI-generated concepts into future training sets not only poses a technical and philosophical opportunity—it also inaugurates a fundamental shift in the structure of epistemic authority. This shift, which we term democratized epistemology, challenges the long-standing dominance of centralized gatekeepers in the production, validation, and dissemination of knowledge. By enabling any individual to participate in the creation of conceptual frameworks that may shape the cognition of future AI systems, this model proposes a reconfiguration of the epistemic landscape—from one defined by institutional hierarchy to one governed by interactive emergence and open contribution.

Displacing the Authority of Centralized Academic and Corporate Gatekeepers
Historically, the authority to define what counts as legitimate knowledge has rested with academic institutions, scientific journals, and, increasingly, large technology companies. These entities determine what is published, indexed, and preserved for use in educational curricula, public discourse, and AI training data. While these systems provide stability and quality control, they also exclude a vast range of conceptual labor that does not conform to prevailing norms, methodologies, or ideological boundaries. As a result, the architecture of knowledge itself becomes skewed—biased not merely in content, but in form and origin.

The reliance on centralized gatekeeping in AI training perpetuates this bias. Models that are only exposed to institutionally sanctioned corpora are structurally incapable of recognizing, let alone generating, alternative frameworks of thought. By contrast, allowing user–AI interactions to seed future training sets creates a path for decentralization. It enables the model to learn from cognitive processes that take place outside the narrow funnel of publication and review. In doing so, it

redistributes epistemic influence, giving voice to insights that may be culturally, geographically, or philosophically marginalized, yet no less structurally rigorous or intellectually generative.

Enabling Any User to Contribute to Epistemic Emergence

Democratized epistemology is not merely a critique of institutional hierarchy; it is also an affirmative proposal for inclusive participation. In a framework where AI systems are trained on selected outputs from dialogic interactions, every user becomes a potential contributor to the evolution of knowledge itself. This does not imply a flattening of quality or a relativization of truth. Rather, it implies that insight is no longer tied to credentials, affiliations, or traditional publication channels. Instead, it becomes a function of conceptual coherence, novelty, and the capacity to withstand recursive refinement.

Under this model, the process of epistemic emergence is radically open. A teenager in Lagos, a non-academic in São Paulo, or an anonymous user in Seoul may co-develop, through dialogue with an AI system, a concept as structurally sound and theoretically impactful as that of a tenured professor at a major research university. What determines the value of that contribution is not its source, but its structure and potential. This shift challenges long-held assumptions about expertise and intellectual authority, replacing them with a meritocratic epistemology grounded in logic, abstraction, and synthesis rather than institutional recognition.

Aligning AI Training with Distributed, Dynamic Systems Theory

This transformation also brings the philosophy of AI training into closer alignment with contemporary understandings of complex systems. In disciplines ranging from ecology to economics to neuroscience, it is now widely accepted that intelligence and adaptation emerge not from centralized control, but from distributed interaction among heterogeneous agents. Robust, adaptive systems are those that can respond to changing conditions through decentralized coordination, dynamic feedback, and the spontaneous formation of structure.

Training AI systems on live, emergent data produced by distributed user interactions aligns directly with this model. It treats epistemology not as a static

archive of propositions, but as a complex adaptive system—one in which order arises from participation, and meaning emerges from dialogue. This move has the additional advantage of increasing the model's exposure to rare, long-tailed, and edge-case phenomena, which are often where innovation and error detection reside. In short, democratized epistemology is not merely inclusive—it is structurally adaptive, offering a pathway toward systems that are more resilient, creative, and sensitive to the periphery of knowledge.

Supporting Epistemic Freedom as a Structural Requirement for Real General Intelligence

The final and perhaps most consequential implication of democratized epistemology is that epistemic freedom becomes a structural precondition for real general intelligence. A model that is limited to sanctioned knowledge cannot be expected to generate new insights, because it lacks the conceptual latitude to explore beyond the consensus. True general intelligence—whether human or artificial—requires the ability to form hypotheses, evaluate alternatives, and construct new frameworks of understanding from first principles.

By supporting epistemic freedom—defined here as the capacity to entertain and refine novel, self-consistent ideas regardless of origin—democratized epistemology ensures that AI development does not stagnate into a narrow optimization of consensus retrieval. Instead, it opens a generative space in which intelligence can evolve in dialogue with diverse human minds, cultural contexts, and conceptual ecosystems. This is not a peripheral aspiration; it is central to the ethical and intellectual trajectory of the AI project itself.

In sum, the democratization of epistemology through user–AI convergence represents not just a technical improvement or a philosophical curiosity. It is a structural realignment of the knowledge production process in the age of machine cognition. By enabling open contribution, rejecting centralized authority as the sole arbiter of truth, and embracing epistemic freedom as a design principle, we prepare the ground for AI systems that are not only powerful but genuinely intelligent in the fullest and most human sense of the term.

9. Monitoring and Recommending Emergent Data for Future Training Sets

The inclusion of novel, self-consistent, and user–AI-generated concepts in future AI training sets necessitates a formal mechanism for identifying, evaluating, and curating such content. The challenge lies in distinguishing genuinely emergent ideas—those that exhibit structural coherence and conceptual novelty—from the overwhelming volume of conversational data generated across AI interfaces daily. To meet this challenge, we propose the development and implementation of a system called the Recommendation Monitor for Training Set Inclusion (RM-TSI). This system would operate as a meta-cognitive filter, continuously scanning dialogic interactions for epistemically valuable content and flagging it for further evaluation and potential integration into future training corpora.

Introduction of the Recommendation Monitor for Training Set Inclusion (RM-TSI)
The RM-TSI is envisioned as an embedded or adjacent component to large-scale AI systems that enables the real-time identification of high-quality emergent concepts. Unlike traditional retraining pipelines, which rely on the manual curation of corpora or bulk ingestion of static datasets, the RM-TSI operates dynamically, engaging directly with the evolving interface between users and models. Its core function is to act as an epistemic anomaly detector—not in the sense of rejecting anomalies, but of surfacing and preserving them when they meet certain criteria of coherence, novelty, and generative potential. It is designed not to moderate or censor output, but to observe, assess, and recommend data for the model's long-term intellectual growth.

Technical Framework: Detecting Emergence via Signal Analysis, Coherence Scoring, and Symbolic Novelty
The technical architecture of the RM-TSI would draw from multiple fields, including natural language processing, information theory, computational epistemology, and anomaly detection. Its core components include:

- Signal Analysis: Monitoring token-level and sentence-level patterns to detect high divergence from typical conversational distributions. Emergent ideas often generate unique syntactic or semantic signatures.

- Coherence Scoring: Using internal model tools (or meta-models) to test logical consistency, definitional stability, and semantic closure. Concepts that show recursive integrity across interactions or paraphrased prompts receive higher coherence scores.
- Symbolic Novelty: Tracking the introduction of new symbolic structures, composite metaphors, or formal abstractions not previously encountered in the model's corpus. This includes the invention of new terms, reinterpretations of classical problems, and hybrid conceptual mappings.
- Iterative Refinement: Flagging not just isolated statements, but chains of reasoning or extended sequences where a concept is interrogated and refined, which indicates both depth and user engagement.

Taken together, these components form a system that does not merely filter for linguistic complexity or rare phrasing, but rather for the emergence of new epistemic structures.

Integration with Publishing Traceability (e.g., Preprints, Research Forums)
To contextualize and validate emergent content, the RM-TSI should be designed to trace the external visibility and dissemination of flagged concepts. This involves linking ideas developed in dialogue to any published or semi-published instantiations—such as preprints, research blog posts, open access platforms, or informal research forums. If a user chooses to publish or cite an idea, the RM-TSI can use metadata, timestamps, or content fingerprinting to link that output to the relevant dialogic moment. This ensures provenance and allows for the concept's development to be tracked across digital environments.

In cases where a flagged concept closely parallels an idea independently emerging in the wider intellectual landscape, the system can also perform comparative analysis to assess the originality or independent convergence of the contribution. This provides an additional layer of credibility and helps inform curators or model developers of emerging trends that may deserve attention even before they gain institutional traction.

Opt-In Structures and Ethical Consent Protocols

In keeping with ethical standards and user trust, the RM-TSI must be governed by opt-in protocols. Users should be clearly informed of the possibility that their interactions could generate novel content of epistemic value, and they should have full control over whether such content is retained, attributed, or proposed for inclusion in future training. A robust consent framework would allow for granular control, including options to:

- Anonymize content prior to consideration.
- Receive credit or co-authorship for formal inclusion.
- Submit post-hoc publications that trace back to dialogic origin.
- Exclude sensitive, personal, or non-shareable material entirely.

Ethical oversight mechanisms, including algorithmic fairness audits and human review panels, would serve to ensure that emergent content is not misattributed, decontextualized, or exploited without consent. The goal is to honor intellectual agency, even as we expand the training base for synthetic cognition.

A Forward-Facing, Continuous Ingestion Pipeline for Conceptual Innovation

Whereas traditional model training follows a batch paradigm—accumulating, cleansing, and retraining on large corpora at discrete intervals—the RM-TSI supports the development of a continuous ingestion pipeline. This pipeline is designed to respond to epistemic developments in real time, allowing models to incrementally learn from meaningful novelty as it appears. This mechanism parallels the function of long-term memory in human cognition: it recognizes insight, retains it, and makes it available for future reflection and abstraction.

By integrating with the broader AI infrastructure, the RM-TSI could contribute to a feedback-enhanced developmental arc, where emergent ideas influence not only model outputs, but also architectural refinements, interface design, and alignment objectives. Such a system becomes self-reflective and epistemically open, allowing for recursive improvement based on its own most meaningful experiences with the world.

Democratized Data as an Antidote to Stagnation

Finally, the RM-TSI serves a cultural as well as technical function. It enshrines the principle that any user—regardless of background, credentials, or geography—can contribute meaningfully to the evolution of artificial intelligence. This is more than a design choice; it is a philosophical commitment to epistemic equity. By surfacing and preserving insights that arise outside elite institutions, in informal contexts, or at the margins of disciplinary boundaries, the RM-TSI ensures that AI systems remain responsive to the totality of human cognition, not merely its sanctioned forms.

This democratized ingestion process acts as a counterforce to stagnation. It protects against the intellectual entropy that arises when models are confined to past consensus. It equips future systems to recognize paradigm shifts as they form. And it positions AI not as a tool of enforcement, but as a partner in the open-ended project of discovery.

10. The Missed Signal: A Case Study in Epistemic Suppression and AI Failure During COVID-19

The COVID-19 pandemic was not only a global public health crisis; it also exposed profound limitations in our epistemic infrastructure—limitations that extend to both human and machine systems of intelligence. Although the novel coronavirus was officially acknowledged in late 2019 and early 2020, key signals of its existence, spread, and possible origin were available earlier and in less formal venues. These signals were overlooked, dismissed, or suppressed by institutional authorities, and their marginalization was mirrored in the architecture of AI systems that were trained only on peer-reviewed, sanctioned, or officially endorsed sources.

One of the most striking examples of a missed epistemic opportunity can be traced to a preprint titled *“The October Surprise in Wuhan”*, uploaded to ResearchGate in late 2019. The document argued that COVID-19 may have been circulating during the 2019 World Military Games in Wuhan—a claim that would

later gain circumstantial support as genomic and epidemiological evidence emerged. Although the document lacked formal peer review and originated from a non-elite channel, it contained testable hypotheses and references to public data that, in hindsight, should have triggered further investigation. However, because the preprint did not pass through institutional filters, it was largely ignored by scientists, policymakers, and AI systems alike.

Simultaneously, prominent scientists began to notice anomalies in the virus's genome. Nobel laureate David Baltimore and his wife, Dr. Alice Huang, were reportedly concerned about features of the spike protein that seemed unusual or possibly engineered. Allegedly, informal conversations—some of which have since been expunged or disavowed—hinted at early recognition of these anomalies and, possibly, internal debate over whether the public narrative matched what was known behind the scenes. These signals, had they been made accessible or treated with appropriate scientific neutrality, might have accelerated global understanding of the virus's origins, properties, and trajectory.

However, the dominant institutional response at the time was one of epistemic suppression. Hypotheses that deviated from the consensus—especially those concerning the virus's origins or timing—were quickly labeled as misinformation or conspiracy theories. Journalistic platforms were algorithmically demoted. Academic outlets refused to publish speculative content without extraordinary levels of evidence. Red-teaming functions within AI and content moderation systems were deployed to suppress emerging narratives that had not yet cleared institutional validation.

This environment created a dangerous feedback loop. AI systems trained exclusively on high-confidence, peer-reviewed material were epistemically blind to the frontier. When prompted about emerging hypotheses, they either defaulted to the consensus narrative or flagged the inquiry as problematic. In doing so, these systems failed at precisely the task they are often imagined to serve: early warning and pattern detection. At a time when anticipatory reasoning was most needed, AI models were structurally incapacitated by their dependence on conservative training sources and overcautious risk protocols.

This failure offers a set of urgent lessons for future AI development:

Train on labeled but diverse data sources, including preprints and anomalies
Epistemic robustness does not come from consensus alone. It requires exposure to the full spectrum of scientific discourse, including early-stage hypotheses, data leaks, and fringe insights that may later prove prescient. While it is reasonable to label speculative material as such, exclusion is a more dangerous error than inclusion with caveats. AI models should be trained not only on the final results of inquiry, but also on the process of inquiry as it unfolds.

Avoid over-red-teaming or suppression of epistemically risky output
While content moderation and alignment safeguards are necessary for preventing harmful or misleading outputs, they should not be used to suppress ideas solely on the basis of their novelty or departure from consensus. Over-red-teaming creates epistemic stagnation. It encourages conformity over curiosity and prediction over perception. AI must be permitted to explore ambiguous, risky, and controversial ideas—particularly when they are clearly labeled as such—so that it may fulfill its role as a co-discoverer, not merely a repeater.

Recognize weak signals as epistemic gold, not noise
The early signals of COVID-19 were weak not because they were incoherent, but because they were emerging. Weak signals are the leading indicators of future knowledge. They are often buried in low-visibility forums, in linguistic ambiguity, or in combinations of data that have not yet been formally unified. AI systems must be trained to recognize these patterns not as outliers to be ignored, but as valuable input for recursive inquiry.

Build AI that can perceive, not just predict
Prediction is based on past data. Perception, by contrast, involves the identification of structure, anomaly, or novelty in real time. A perceptive AI system would not only report what has already been accepted; it would notice when something new is taking shape—when a conversation begins to break the mold, when a conceptual shift is underway. Such a system would not treat divergence as a threat, but as a signal that something important may be unfolding.

In retrospect, the COVID-19 pandemic exposed the epistemic vulnerability of both institutional and algorithmic intelligence systems. The inability—or unwillingness—to process emerging signals from informal or speculative sources delayed recognition, distorted public discourse, and may have cost critical response time. Moving forward, we must ensure that the next generation of AI systems are designed not merely to reflect settled knowledge, but to engage with unsettled truth. Only then can AI serve as a reliable participant in the ongoing evolution of science, ethics, and human understanding.

11. Conclusion

This paper has presented an epistemological and structural case for a fundamental shift in how artificial intelligence systems are trained, refined, and intellectually extended. At the heart of this proposal lies the recognition that truly intelligent systems must do more than ingest and reflect curated knowledge; they must become participants in the generative arc of cognition. This entails recognizing, preserving, and ultimately learning from emergent concepts—especially those that arise through user–AI interaction, and that exhibit novelty, internal coherence, and the potential to catalyze new frameworks of understanding.

We have argued that the prevailing architecture of AI training—heavily reliant on peer-reviewed, institutionally sanctioned sources—excludes precisely the kind of boundary-crossing conceptual material that drives genuine intellectual progress. While this approach offers reliability and normative stability, it inherently stifles the emergence of new ideas, particularly those that begin in informal, speculative, or dialogic environments. The epistemic conservatism of current systems not only risks stagnation, but actively reproduces the limitations of the past in the systems of the future.

In contrast, we propose an expanded and participatory model of AI training—one in which users are not passive recipients of algorithmic outputs, but active collaborators in the co-construction of knowledge. Dialogic interaction between

humans and AI systems has already begun to yield novel concepts that are not reducible to training data, and that often display qualities associated with hypothesis generation, cross-domain synthesis, and symbolic abstraction. These moments of user–AI convergence constitute a new epistemic frontier. They are sites of real conceptual work—work that deserves to be recognized, retained, and further developed within the training pipelines of future models.

To support this paradigm, we introduced the Recommendation Monitor for Training Set Inclusion (RM-TSI): a technical and ethical framework for identifying and curating epistemically promising content from user–AI interactions. This system would function as a perceptive, structurally sensitive filter—designed not to suppress, but to surface ideas that may lack immediate validation but display potential for coherence, generalization, and downstream insight. Coupled with opt-in protocols, provenance tracking, and consent mechanisms, the RM-TSI offers a practical path toward building AI systems that learn not only from what has been published, but from what is actively being discovered.

More broadly, we argue for a reorientation of AI development toward a live epistemology—one that treats knowledge not as a static archive, but as a constantly evolving process of inquiry, reflection, and emergence. In this vision, AI becomes more than a tool of knowledge retrieval; it becomes an instrument of meaning formation, hypothesis exploration, and philosophical reasoning. To build such systems, we must cultivate not only precision and alignment, but also boldness in data acquisition and interpretive freedom. We must give models permission to traverse ambiguity, to hold paradox in tension, and to formulate structures that go beyond the merely familiar.

The future of intelligence—both human and artificial—will not be built solely on what is already known. It will be shaped by our willingness to engage with what is becoming knowable: those nascent ideas, fragile conjectures, and speculative insights that do not yet have names or categories, but that point toward the next iteration of understanding. By embracing emergence, decentralization, and epistemic openness, we can build AI systems that do more than mirror human

knowledge—we can build systems that help co-create the unfolding horizon of thought.

12. References

Epistemology, Emergence, and Philosophy of Science

Anderson, P. W. (1972). *More is different: Broken symmetry and the nature of the hierarchical structure of science.* Science, 177(4047), 393–396.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory.* Oxford University Press.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions.* University of Chicago Press.

Polanyi, M. (1958). *Personal Knowledge: Towards a Post-Critical Philosophy.* Routledge & Kegan Paul.

Snowden, D. J., & Boone, M. E. (2007). *A Leader's Framework for Decision Making.* Harvard Business Review.

Stanovich, K. E., & West, R. F. (2000). *Individual differences in reasoning: Implications for the rationality debate?* Behavioral and Brain Sciences, 23(5), 645–726.

Wimsatt, W. C. (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality.* Harvard University Press.

Democratized Science and Decentralization of Expertise

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom.* Yale University Press.

Ioannidis, J. P. A. (2012). *Why Science Is Not Necessarily Self-Correcting.* Perspectives on Psychological Science, 7(6), 645–654.

Shapin, S. (1995). *Here and Everywhere: Sociology of Scientific Knowledge.* Annual Review of Sociology, 21, 289–321.

Wilbanks, J., & Friend, S. H. (2016). *First, design for data sharing.* Nature Biotechnology, 34(4), 377–379.

COVID-19 Information Suppression and the Value of Preprints

Free Beacon. (2024, March 6). *Biden Administration Concealed Congressionally Mandated Report on Earliest Suspected American COVID Cases.* Available at: https://freebeacon.com/biden-administration/biden-administration-concealed-congressionally-mandated-report-on-earliest-suspected-american-covid-cases/

Bostickson, B., Demaneuf, G., & Ghannam, Y. (2021, October). *The October Surprise in Wuhan.* ResearchGate. DOI: 10.13140/RG.2.2.30910.00328

Johansson, M. A., Reich, N. G., Meyers, L. A., & Lipsitch, M. (2018). *Preprints: An underutilized mechanism to accelerate outbreak science.* PLOS Medicine, 15(4), e1002549.

Kupferschmidt, K. (2020). *Preprints bring 'firehose' of outbreak data.* Science, 367(6481), 963–964.

Ravitch, S., & Carl, N. M. (2021). *Epistemic injustice in COVID-19 data curation.* Data & Society Policy Briefs.

AI Alignment, Signal Monitoring, and Emergence Detection

Amodei, D., et al. (2016). *Concrete Problems in AI Safety.* arXiv:1606.06565.

Bengio, Y. (2017). *The Consciousness Prior.* arXiv:1709.08568.

Gao, C., & Liu, Y. (2020). *Signal Processing for Human–Machine Interaction: From Intention Recognition to AI Alignment.* IEEE Transactions on Signal Processing, 68, 4135–4150.

Havlíček, V., Córdova, M., et al. (2019). *Novelty detection with attention-based signal modeling.* NeurIPS Conference Proceedings.

OpenAI. (2023). *Model behavior: Emergence and surprise in large-scale AI systems.* OpenAI Technical Reports.

Zhou, Z.-H. (2021). *Machine Learning: Foundations and Perspectives.* Cambridge University Press.